OrphHCA Documentation

Release 0.5

Tristan Bitard-Feildel

June 10, 2015

Contents

CHAPTER 1

Introduction

1.1 Protein domain

A protein domain corresponds to a conserved region of a protein sequence. Depending on the domain ressouces used, a domain is either first defined based on structural information, followed by a search for similar sequences corresponding to the limits given by the structure, or based only on sequence conservation deduced from similarity searches.

A protein domain can be alone on a protein sequence or can be coupled to other ones to form a particular domain arrangement, i.e. a succession of the same or of different domains along the protein sequence. The comparison of protein domain arrangements can give deep insight into our understanding of: protein evolution, phylogeny relationships between species, protein function, ...

1.2 Sequence annotation

Protein domain annotation methodologies typically use a protein domain database and scan a query proteome against all the models present in the database. The domains are represented inside the database as `Hidden Markov Models` (HMMs). These **HMMs** are built from `Multiple Sequence Alignments` (MSAs) of sequences of protein segments that are classified as belonging to the same domain family.

One of the major difficulty relies on the creation of the domain family set of sequences. As mentioned above, a search for regions sharing similarities between sequences is performed. Families with domains present in a sufficient large number of species will be detected without too much difficulties. However recent domains, domains present only in a specific clade for which too few species are available, or fast divergent protein domain families will usually be missed by methods based only on sequence similarity searches.

1.3 The HCA method

The `Hydrophobic Cluster Analysis` (HCA) *[CG1987] [IC1997]* of protein sequences is a methodology that performs a coupled physico-chemical and topological analysis of the amino acids present on a protein sequence. In globular proteins, the hydrophobic amino-acids present on the regular secondary structures (alpha helices and beta strands) display a typical binary pattern of alternating hydrophobic and non-hydrophobic amino acids, that corresponds to the general trend of hydrophobic residues to be buried inside the protein cores *[JH2003] [RE2007]*. The use of a bidimensional support to represent the protein sequences brings an additional dimension to the binary pattern definition, leading to the definition of constrained binary patterns or *hydrophobic clusters*, through the use of a connectivity distance separating them into distinct units. Positions of *hydrophobic clusters* mainly correspond to those of the regular secondary structures, and can be used to characterize in different ways the protein fold characteristics.

SegHCA *[FG2013]* is a tool based on the **HCA** methodology allowing the detection of high densities of hydrophobic clusters on protein sequences. These hot spots can then be used as a proxy to protein area with a propensity to fold, i.e. protein domains. These areas are called *HCA-segments*.

1.4 OrphHCA

The OrphHCA software has been designed to propose a solution for finding: recent domains, fast diverging domains, or domains on proteomes of clades with only a few number of species. The methodology has been previously tested on a set of *Drosophila* orthologous proteins *[TBF2015]* and was able to detect recent and fast diverging domains.

The workflow of the methodology is presented below:

The methodology can be separated into two steps. The first step, mandatory, corresponds to the domain annotation. `SegHCA` is used to delineate *HCA-segments*, and optionally an annotation with other databases can be performed using `hmmscan`. The annotation is followed by several filtering procedures to detect the conserved *HCA-segments*.

The second step corresponds to a filtering step, during which the generated *HCA-segments* are compared to some other databases or to each others.

1.5 References

Read the Tutorial for a quick start on how to use OrphHCA!

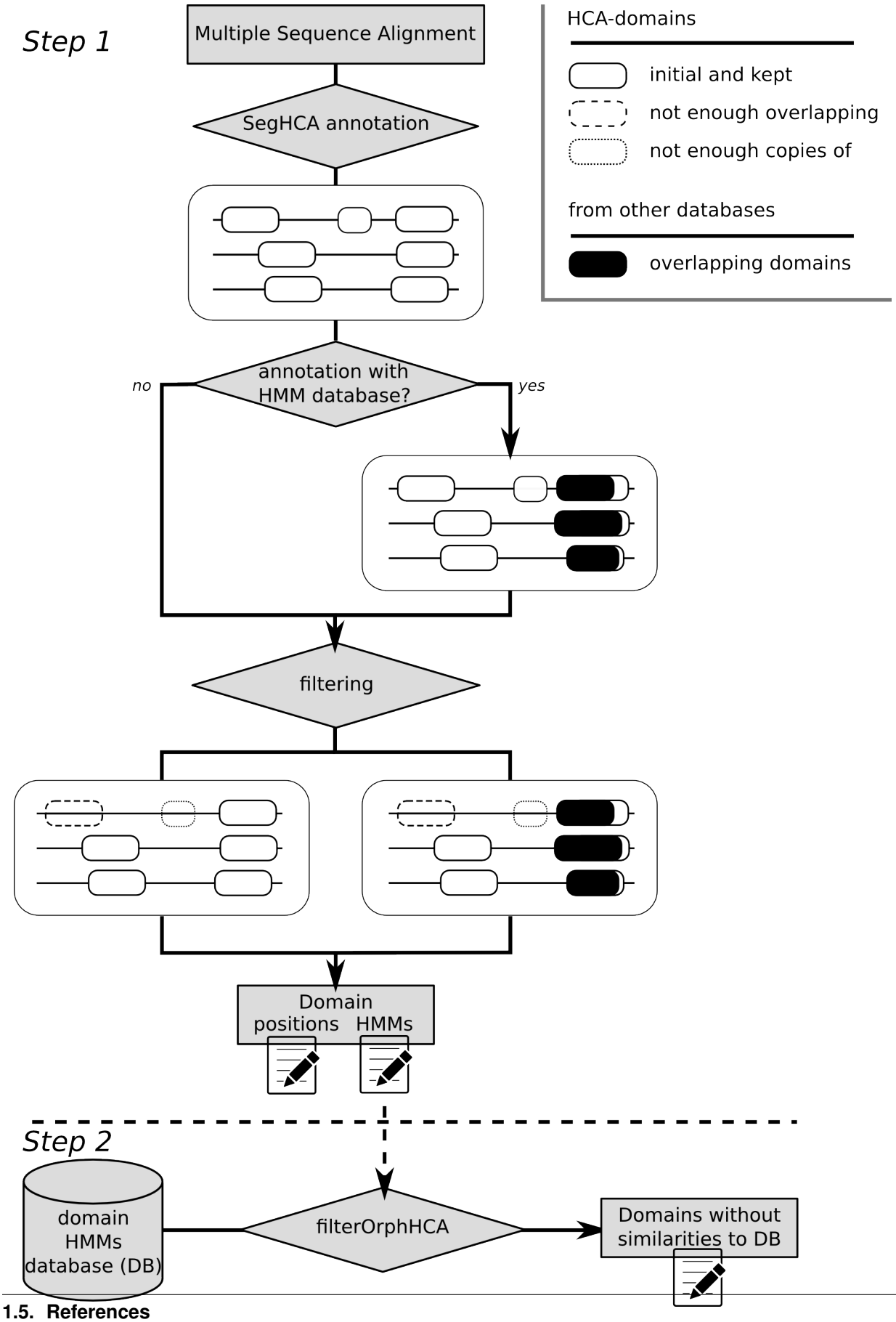

Fig. 1.1: **The OrphHCA workflow.**

CHAPTER 2

Installation

2.1 Requirements

2.1.1 Python requirement

OrphHCA requires Biopython and can be downloaded here.

2.1.2 System requirement

The OrphHCA softwares are using several tools that need to be installed independtly on your computer. The localisation of thee tools have to be specified in the configuration file `PATH.ini`. The `PATH.ini` file is a basic configuration file used by the python ConfigParser module .

> `PATH.ini` syntax:
> [HMM]
> hhsearch:/opt/global/bin/hhsearch
> hmmscan:hmmscan
> ...

The environment variable `ORPHHCA_DATA` toward `PATH.ini` need to be set up.

```
export ORPHHCA_DATA=`pwd`
```

The following executable should be installed on your computer and present in the `PATH.ini` file:

- hhsearch, binary from the hhpred package
- hhblits, binary from the hhpred package
- hhmaker, binary from the hhpred package
- reformat.pl, script from the hhpred package
- hmmscan, binary from the HMMER package
- hmmbuild, binary from the HMMER package
- hmmpress, binary from the HMMER package

An example of `PATH.ini` can be found here.

2.2 Install

The easiest way to install OrphHCA is to use pip:

```
pip install orphHCA --user
```

You can also clone the sources from git and install locally:

```
git clone ssh://git@ebbgit.uni-muenster.de:62246/tbitardfeildel/orphhca.git
cd orphhca.git*
python setup.py install --user
```

alternatively you can download the sources as a zip:

```
http://www.bornberglab.org/pages/orphhca/
http://ebbgit.uni-muenster.de/tbitardfeildel/orphhca/
```

Read the Tutorial for a quick start on how to use OrphHCA!

CHAPTER 3

Tutorial

3.1 Quick description

A more complete description of the method can be found in the Introduction.

OrphHCA is designed to detect conserved hydrophobic segments (called *HCA-segments*) on multiple sequence alignment (MSA).

The input of OrphHCA is a MSA fasta file. OrphHCA is actually distributed as two scripts. The main script **orphHCA** and an utilitary script **filterOrphHCA**.

The **orphHCA** script performs the external domain annotation and the HCA-segments search. Then, it selects the segments corresponding to domains based on their overlaps and their conservation in the MSA. Finally, the script produces a flat file with the domain positions and an hmm database file built with **hmmbuild**.

The **filterOrphHCA** script can be used to compare the created hidden markov models (HMMs) with models from other databases. The script uses the **hhsearch** tool to perform the comparison.

3.2 Getting started

First you will need to install OrphHCA. A complete documentation on how to install OrphHCA can be found in Installation.

Warning: As OrphHCA built the amino-acid sequences from the sequences of the MSA, non amino-acids characters ["*", "!", ".", "?", "-"] are removed. Other characters in the sequences are kept.

3.2.1 Example file

The example file to run orphHCA can be found in the example in the git repository.

3.2.2 Running `orphHCA`

Running orphHCA without specific parameters.

```
$ orphHCA -i examples/EOG7CPB12.fasta -o examples/EOG7CPB12 -w examples/EOG7CPB12/ -v --keep-fas
```

Two files are created: `examples/EOG7CPB12.out` and `examples/EOG7CPB12.hmm`.

The first file ''examples/EOG7CPB12.out" contains the list of domains found in each protein. The format of the file follows the `xdom` syntax.

```
>FBgn0179134_Dsec_1 772
13 61 orph_0 Nan # 12 65
203 258 orph_1 Nan # 202 258
288 328 orph_2 Nan # 287 328
395 772 orph_3 Nan # 388 772
>FBgn0241472_Dyak_1 780
13 61 orph_0 Nan # 12 65
...
```

Each protein entry starts with a fasta header correspoding to the name of the protein sequence, for example `FBgn0179134_Dsec_1` followed by a space character and the length of the protein sequence, here `772` for the protein `FBgn0179134_Dsec_1`.

The lines following the fasta header correspond to domain positions. The line `13 61 orph_0 Nan # 12 65` is made of four required columns, `13 61 orph_0 Nan`, and followed by two commented columns, `12 65`. The numbers `13 61` in the required columns correspond to the start and stop positions of the domain, the position are inclusive and the first amino-acid of the sequence starts at 1. The name `orph_0` corresponds to the domain name and can be shared between the proteins, the `Nan` correspond to the e-value field of the `xdom` and should be ignored as no e-values are computed. The two optional columns `12 65` correspond to the full length of the domain.

The final positions, `13 61`, are computed based on the domain position conservation between the sequences and the original HCA-domain annotation of the protein sequences can be longer, `12 65` in this example. As a matter of comparison the positions `13 61` can be seen as the alignment position, `ali` columns, of the annotation produced by **`hmmscan`** and the columns `12 65` as the envelop of the domain, `env` columns in **`hmmscan`** results.

The second file `examples/EOG7CPB12.hmm` is an hmm file generated from **`hmmbuilt`**. All the domain models are concatened in this file.

3.2.3 Running `filterOrphHCA`

Running filterOrphHCA:

```
$ filterOrphHCA -f examples/EOG7CPB12/kept_fasta/ -i examples/EOG7CPB12.hmm -w examples/filtering_EOG
```

The program takes as an input the directory of the fasta files corresponding to the previously created HMMs, `examples/EOG7CPB12/kept_fasta/`, with an HMM databases corresponding to the fasta file, `examples/EOG7CPB12.hmm`, a working directory, `examples/filtering_EOG7CPB12/` and an external database against which the created models are compared, `pfamA_v27.0_22Oct13.hhm`.

The output file, `examples/EOG7CPB12.filtered.dat` is a tab delineated flat file of four columns.

```
model_name_1 target_name_1 similarity database_of_the_target_1
model_name_1 target_name_2 similarity database_of_the_target_2
...
model_name_2 target_name_1 similarity database_of_the_target_1
...
```

All the targets having a similarity score strictly above the cutoff parameter, `-c 89`, are reported.

3.3 Parameters of `orphHCA`

3.3.1 Required parameters

-i, --input : FILE
the MSA input file

-o, --output : FILE PREFIX
output file prefix (<output>.out : list of domains, <output>.hmm : hmmdatabase)

-w, --workdir : DIR
working directory

3.3.2 Optional parameters

-d, --database
list of the domain hmm databases to use

-s, --seqdb
path to the sequence database used for enrichment

-c, --core
number of cores to use; default=1

--perc-hca
minimal percentage of sequences in the MSA that should have a domain , default=20

--nb-hca
minimal number of sequences in the MSA that should have a domain

--perc-over, default=80
minimal percentage of overlap allowed between hca segment for them to be considered as part of the same domain

--nb-over
minimal number of overlapping amino-acids between two hca segments to consider them as the same

--hca-size
minimal size to consider a hca segment as a domain, default=30

--perc-hmm
maximal percentage of overlap allowed between a hca segment and a hmm domain , default=0

--nb-hmm
maximal number of overlapping amino-acids allowed between an hca segment and an hmm domain

--keep-fas
keep fasta results, fasta alignment are needed by hhsearch in the filtering program

-v, --verbose
active/inactive verbose mode

3.4 Parameters of `filteringOrphHCA`

3.4.1 Required parameters

`-f, --fastadir`
the directory with fasta alignments

`-i, --inputfile`
the hmm database corresponding to the fasta alignments

`-w, --workdir`
the working directory

`-d, --database`
the list of hmm database to which the fasta alignments are compared to

`-o, --output`
the list of model that are similar to an other model in a database

`-c, --cutoff`
the similarity cutoff

3.4.2 Optional parameters

`-v, --verbose`
activate verbose mode

`-h, --help`
show this help message and exit

CHAPTER 4

Glossary

HCA-segment An HCA-segment corresponds to a high density area of hydrophobic clusters

hydrophobic cluster An hydrophobic cluster is defined as a succession of strong hydrophobic residues separated by less than a given distance in amino acids, called connectivity distance. The strong hydrophobic residues are V, I, L, M, F, Y, W and the connectivity distance is 4 in the standard use of HCA approach, in which the α-helix is used as a two-dimensional support for the 2D HCA transposition of the sequence. These parameters provided the best correspondance between the positions of clusters and regular secondary structures (*[SW1992]*, *[IC1997]*).

4.1 References

Bibliography

[CG1987] Gaboriaud C, Bissery V, Benchetrit T, Mornon JP. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett. 1987 Nov 16;224(1):149-55.

[IC1997] Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci. 1997 Aug;53(8):621-45.

[JH2003] Hennetin J, Le Tuan K, Canard L, Colloc'h N, Mornon JP, Callebaut I. Non-intertwined binary patterns of hydrophobic/nonhydrophobic amino acids are considerably better markers of regular secondary structures than nonconstrained patterns. Proteins. 2003 May 1;51(2):236-44.

[RE2007] Eudes R, Le Tuan K, Delettré J, Mornon JP, Callebaut I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. BMC Struct Biol. 2007 Jan 8;7:2.

[FG2013] Faure G, Callebaut I. Comprehensive repertoire of foldable regions within whole genomes. PLoS Comput Biol. 2013 Oct;9(10):e1003280

[TBF2015] Bitard-Feildel T, Heberlein M, Bornberg-Bauer Erich and Callebaut I. Detection of Orphan Domains in Drosophila using "Hydrophobic Cluster Analysis" Biochimie *accepted*

[IC1997] Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci. 1997 Aug;53(8):621-45.

[SW1992] Woodcock S, Mornon JP, Henrissat B. Detection of secondary structure elements in proteins by hydrophobic cluster analysis Protein Eng. 1992 5(7):629-635.

Symbols

C

H